Least Squares Estimation of Probability Measures in the Prohorov Metric Framework

H.T. Banks and W. Clayton Thompson

Center for Research in Scientific Computation

Center for Quantitative Sciences in Biomedicine

North Carolina State University, Raleigh, NC 27695-8212

November 30, 2012

Abstract

We consider nonparametric estimation of probability measures for parameters in problems where only aggregate (population level) data are available. We summarize an existing computational method for the estimation problem which has been developed over the past several decades [4, 8, 16, 21]. Significant new theoretical results are presented which establish the existence and consistency of very general (ordinary, generalized and other) least squares estimates for the measure estimation problem.

Mathematics Subject Classification:62G07,34A55,46S50,93E24.

Key words: least squares inverse problems for probability measures; Prohorov metric; nonparametric estimation; existence and consistency of estimators.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **30 NOV 2012**	2. REPORT TYPE	3. DATES COVERED **00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE **Least Squares Estimation of Probability Measures in the Prohorov Metric Framework**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **North Carolina State University,Center for Research in Scientific Computation,Center for Quantitative Sciences in Biomedicine,Raleigh,NC,27695-8212**	8. PERFORMING ORGANIZATION REPORT NUMBER **CRSC-TR12-21**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
We consider nonparametric estimation of probability measures for parameters in problems where only aggregate (population level) data are available. We summarize an existing computational method for the estimation problem which has been developed over the past several decades [4, 8, 16, 21]. Significant new theoretical results are presented which establish the existence and consistency of very general (ordinary, generalized and other) least squares estimates for the measure estimation problem.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**31**	

1 Motivation

In a standard nonlinear regression problem, a mathematical model is proposed which links one or more states of interest to the independent variables (regressors) of an experiment and to a vector of parameters whose values are unknown to the experimenter. An experiment is then conducted on the physical or biological system and data is collected for one or more states of interest. The unknown parameters of interest are then estimated in an *inverse* or *parameter estimation* problem, the theory for which is well-established [15, 28, 31]. Yet in many situations physical, biological, or experimental limitations do not permit one to sample individual data directly. Rather, one obtains data at the *aggregate* level as multiple individuals are sampled. It is commonly assumed that the states of interest for these individuals are described by a single mathematical framework, but that each individual is described by a unique set of parameters within that framework. For instance, the growth of mosquitofish [9, 16, 17] and shrimp [5, 11, 13] have been shown to be described by a size-structured partial differential equation model in which the rate of individual growth is assumed to vary probabilistically across the population. HIV replication data has been shown to be accurately described by a cellular-level model in which intracellular delays vary from cell to cell [7]. The probabilistic distribution of parameters has also been observed in models of electromagnetic polarization [8, 10, 18, 19] and in the deformation of viscoelastic materials [22]. These examples are considered at greater length in the recent book [2].

In each of these examples, one has a mathematical model which describes the behavior of an individual but data which has been sampled from an entire population of individuals. Thus, in the context of the mathematical model, one has information not on the value of a fixed, single parameter, but rather on the distribution of parameters which characterizes the behavior of the entire population. It is this probability distribution which one seeks to estimate. Significantly, the data is sampled from the state space of the mathematical system and not from the parameter space; thus one *does not sample directly from the distribution of interest.* In developing a framework for this estimation problem, one encounters a rich body of mathematical theory. In this document, we summarize a computational method for the estimation problem which has been developed and tested computationally over the past several decades [4, 8, 16, 21] (Section 5). In Section 4 below, significant new results are given which establish the existence and consistency of the least squares estimator for this nonparametric estimation problem. First, we formally define the estimation problem.

Suppose that the quantities of interest for a single individual can be described by the mathematical model

$$\begin{aligned}\frac{dy}{dt} &= g(t, y(t); q, \psi),\\ y(t_0) &= y_0. \qquad (1.1)\end{aligned}$$

The parameter vector $q \in \mathbb{R}^r$ is specific to each individual within the population while the parameter vector $\psi \in \mathbb{R}^u$ describes parameters common to all individuals within the population (e.g., environmental factors). The observation model solution is given by

$$y(t; \theta, \psi) = Cf(t; q, y_0, \psi) \qquad (1.2)$$

where $\theta = (q, y_0) \in \mathbb{R}^{r+s} = \mathbb{R}^p$. It is assumed $f(t; \theta, \psi) \in \mathbb{R}^s$ and $C \in \mathbb{R}^{l \times s}$ so that $y \in \mathbb{R}^l$. (In the notation that follows, we tacitly assume $l = 1$; this is only for convenience and all theory presented holds for vector observations.) It is assumed that $\theta \in \Theta$ and $\psi \in \Psi$ for all individuals in the population, where Θ and Ψ are sets of admissible parameters.

For the aggregate data problem, one can consider n observations as random variables resulting from the direct sampling of the mean population state, but measured subject to random error. Then it is possible to define the random variables

$$V_j = v(t; P_0, \psi_0) + \mathcal{E}_j \tag{1.3}$$

for $j = 1, \ldots, n$ where

$$v(t; P, \psi) = E[Cy(t; \cdot, \psi)|P] = \int_\Theta Cy(t; \theta, \psi) dP(\theta),$$

and the random variables $\mathcal{E}_j$ represent measurement noise, modeling error, microfluctuations, etc. Let $\vec{\mathcal{E}} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)$. It is assumed that the first two central moments of the random vector $\vec{\mathcal{E}}$ are

$$\begin{aligned} E[\vec{\mathcal{E}}] &= \vec{0} \\ Var[\vec{\mathcal{E}}] &= R. \end{aligned} \tag{1.4}$$

It is most commonly assumed that the random variables $\mathcal{E}_j$ are independent and identically distributed, so that $R = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix. In the theory presented in this document, we typically make this assumption of a *constant* or *absolute* error model. However, this is not strictly necessary and the results presented can be generalized to include a wide array of statistical models which are encountered in practical problems. Extensions to other error models are considered in Section 6.

Given n realizations v_j of the random variables V_j (which we will sometimes write $\vec{v}$ and $\vec{V}$ for notational convenience), the goal of an inverse or parameter estimation problem is to produce an estimate of the hypothetical true parameters P_0 and ψ_0. Of course, the estimated parameters should be those that best fit the data in some appropriate sense. Thus this problem first involves a choice of framework in which to work. Given that choice of framework, one must establish a set of theoretical and computation tools with which to treat the parameter estimation problem. For the results presented here, we focus on a frequentist approach using general least squares estimation. Theoretical results for likelihood estimation (also in a frequentist framework) can be established with little difficulty from the results presented here. For the moment, we do not consider a Bayesian approach to the estimation of the unknown distribution P_0. There does seem to be some commonality between the nonparametric estimation of a probability distribution and the determination of a Bayesian posterior estimator [21]. However to our knowledge, a comprehensive comparison of the two methods (either theoretical or computational) has not been performed.

We remark here that the estimation of the incidental parameter ψ is not of primary interest in this document. Techniques for the estimation of ψ fall entirely within the theory of classical nonlinear least squares. The parameter is included in the formulation above to provide clear indication that the theory presented below for the nonparametric estimation of a probability distribution is compatible with the simultaneous estimation of an incidental parameter. For instance, in Equation (1.5) below, one can define the estimator (P_n, ψ_n) in $(\mathcal{P}(\Theta), \Psi)$. Without loss of generality, it will be assumed that ψ is known.

For the least squares problem, define the estimator

$$P_n = \arg \min_{P \in \mathcal{P}(\Theta)} J_n(\vec{V}, P) = \arg \min_{P \in \mathcal{P}(\Theta)} \sum_{j=1}^{n} \left(V_j - v(t_j; P)\right)^2 . \tag{1.5}$$

We remark that P_n is itself a random variable in that it is a function of the random variables V_j (and hence $\mathcal{E}_j$). This dependence is generally suppressed with the exception of the subscripted n, but should be carefully noted, particular in the consideration of the existence and consistency of the estimator (Section 4). The inverse problem is then to use realizations v_j of the random variables V_j to compute

$$\hat{P}_n = \arg \min_{P \in \mathcal{P}(\Theta)} J_n(\vec{v}, P) = \arg \min_{P \in \mathcal{P}(\Theta)} \sum_{j=1}^{n} \left(v_j - v(t_j; P)\right)^2 . \tag{1.6}$$

However, one cannot typically compute $\hat{P}_n$ as defined. In most practical problems, the model $v(t; P, \psi)$ cannot be computed exactly and must be approximated with $v^N(t; P, \psi)$ by some numerical scheme (e.g., finite difference methods, Galerkin methods, etc.). Similarly, the space $\mathcal{P}(\Theta)$ has (uncountably) infinitely many elements so that it must also be approximated by some computationally tractable sets $\mathcal{P}_M(\Theta)$. Thus, given a set of realizations $\{v_j\}$ of the random variables V_j, what one computes in practice is

$$\hat{P}_{n,M}^N = \arg \min_{P \in \mathcal{P}_\mathrm{M}(\Theta)} J_n^N(\vec{v}, P) = \arg \min_{P \in \mathcal{P}_\mathrm{M}(\Theta)} \sum_{j=1}^{n} \left(v_j - v^N(t_j; P)\right)^2 . \tag{1.7}$$

The immediate question of interest is how these formal definitions relate back to the actual quantity of interest, the unknown 'true' probability measure P_0. In considering this question, we see that several additional questions must be answered. First, it must be shown that the least squares estimator P_n given by (1.5) is well-defined. The next question is computational: as M and N grow large, is it necessarily true that $\hat{P}_{n,M}^N$ converges (in some sense) to $\hat{P}_n$? Of course, the answer to this question depends largely upon the approximation schemes used. For instance, one could define $\mathcal{P}_M(\Theta)$ to be the subset of the space of probability measures consisting of those measures with a specific parametric form. While this technique has the advantage of creating a standard nonlinear estimation problem, it may lead to inaccurate and misleading results unless there is strong evidence to suggest a particular parametric form for the unknown measure. In this document, we are concerned with *nonparametric estimation*, so that only a minimal set of restrictions is placed on the class of admissible measures.

The remaining question is statistical. Assuming that $\hat{P}_{n,M}^N$ approaches $\hat{P}_n$ as M and N grow large, how does this estimate compare with P_0? Put another way, given any fixed n observations, one obtains an estimate $\hat{P}_n$ of P_n. How does this estimate improve as more data is collected (that is, as n grows large). This is a question of the *consistency* of the least squares estimator P_n.

As will be shown there is a natural setting, which we will call the *Prohorov Metric Framework* (PMF), in which these questions can be answered for parameter estimation problems such as these, in which the unknown parameter is a probability distribution. We begin by describing the Prohorov metric on the space of probability measures and derive some properties which will be useful in answering the questions posed above. Under a fairly general set of conditions, in is shown that the estimator P_n is well-defined (in the sense that it is a measurable function which maps the space of data to the space of probability measures). Next, this estimator is shown to be consistent, and conditions for computational approximation and convergence are given. Finally, the statistical model (1.3) is revisited and the results of this document are extended to a larger class of problems.

2 The Prohorov Metric

We begin with several general definitions and theorems which are meant to motivate the PMF and provide some background. No proofs are given for this motivating material, although references are provided. Details proofs are provided for the more interesting features of the PMF. A number of the results presented can be founded scattered through existing literature. Many of the results of the next two sections (and some alternative proofs) can be founded in [25, 33] and have been usefully organized into an easy-to-read series of notes [27].

First, the Riesz Representation Theorem on the space of bounded continuous functions is stated. This theorem can be used to characterize the weak* topology on the continuous dual of the space of bounded continuous functions, which provides an intuitive motivation for the weak topology on the space of probability measures. It is no surprise then that the two topologies are equivalent on the space of probability measures. Next the Prohorov metric is defined and is shown to metrize the weak topology of measures. The Prohorov metric is then used to establish several desirable properties of the space of probability measures.

Consider the metric space Θ with its metric d, which we can write together as (Θ, d). Define the space $C_B(\Theta) = \{f : \Theta \to \mathbb{R} | f \text{ bounded, continuous}\}$.

Theorem 2.1 (Riesz). *Assume (Θ, d) is a compact (Hausdorff[1]) space. For every $f^* \in C_B(\Theta)^*$ (the continuous dual of the space $C_B(\Theta)$), there exists a unique finite signed Borel measure μ such that*

$$f^*(f) = \int_\Theta f(\theta) d\mu(\theta)$$

for all $f \in C_B(\Theta)$. Moreover, $||f|| = |\mu|(\Theta)$.

Proof. See [30, pg. 357-358]. □

Given this identification, we may write $f^* = f^*_\mu$ when convenient. We see that the set $\mathcal{P}(\Theta)$ of probability measures on (Θ, d) can be identified with those $f^*_\mu \subset C_B(\Theta)^*$ such that $f^*_\mu(f) \geq 0$ for all $f \in C_B(\Theta)$ and $\left|\left|f^*_\mu\right|\right| = \mu(\Theta) = 1$. Thus we have, in a sense, that $\mathcal{P}(\Theta) \subset C_B(\Theta)^*$. In fact, given any $f \in C_B(\Theta)$, the map from $C_B(\theta)$ into $\mathbb{R}$ given by $f^{**}_f(f^*) = f^*(f)$ defines the

[1] The assumption that Θ is Hausdorff will be maintained throughout this document.

natural embedding of $C_B(\Theta) \hookrightarrow C_B(\Theta)^{**}$. The image of f^{**} induces a topology on the space $C_B(\Theta)^*$, known to functional analysts as the weak* topology [2, 49–57]. (That is, $f_n^* \xrightarrow{w^*} f^*$ if and only if $f_n^*(f) \to f^*(f)$ for all $f \in C_B(\Theta)$.) When viewed in the context of $\mathcal{P}(\Theta) \subset C_B(\Theta)^*$, this is the *weak convergence of measures* known from the theory of probability and stochastic processes.

With this motivation, we now turn to the problem of characterizing the weak topology of measures.

Definition 2.2. *Let (Θ, d) be any metric space (not necessarily compact) and define the set $C_B(\Theta)$ as above. Given any probability measure $P \in \mathcal{P}(\Theta)$ and some $\epsilon > 0$, an ϵ-neighborhood of P is*

$$B_\epsilon(P) = \left\{ Q \middle| \left| \int_\Theta f(\theta) dQ(\theta) - \int_\Theta f(\theta) dP(\theta) \right| < \epsilon, \text{ for all } f \in C_B(\Theta) \right\}. \tag{2.1}$$

Comparing the Riesz Represention Theorem (Theorem 2.1) with the definition of $B_\epsilon(P)$, there is a clear connection between the open balls on $\mathcal{P}(\Theta)$ and the weak topology of measures. In fact, we may take the collection of all open balls as the *definition* of the weak topology of measures [25, pg. 236]. Alternatively, we have the following equivalent characterizations of the weak topology.

Theorem 2.3. *Let Θ be a topological space with σ-algebra Σ_Θ. Let $P \in \mathcal{P}(\Theta)$. The following are equivalent:*

1. $B_\epsilon(P)$;
2. $\{Q | Q(C) < Q(C) + \epsilon, C \subset \Theta \text{ closed}\}$;
3. $\{Q | Q(O) < Q(O) + \epsilon, O \subset \Theta \text{ open}\}$;
4. $\{Q | Q(F) < Q(F) + \epsilon, F \in \Sigma_\Theta, P(\partial F) = 0 \text{ (such sets are called P-continuity sets)}\}$.

Proof. See [25, pgs. 236-237]. □

The weak topology of measures, in turn, gives rise to notions of weak (topological) convergence of measures.

Definition 2.4. *Given a sequence of measures $P_M \in \mathcal{P}(\Theta)$ for all $M = 1, \ldots, \infty$, we say P_M converges weakly to P, $P_M \xrightarrow{w^*} P$, if any one (and hence all) of the following equivalent conditions holds:*

1. $\left|\int_\Theta f(\theta) dP_M(\theta) - \int_\Theta f(\theta) dP(\theta)\right| \to 0$ *for all* $f \in C_B(\Theta)$;
2. $\limsup P_M(C) \le P(C)$ *for all C closed in Θ*;
3. $\liminf P_M(O) \ge P(O)$ *for all O open in Θ*;
4. $\lim P_M(F) = P(F)$ *for all sets $F \in \Sigma_\Theta$ such that F is a P-continuity set.*

The equivalence of the above notions of convergence is often referred to as the portmanteau theorem [25, pgs. 11-12]. We remark that the notation $P_M \xrightarrow{w^*} P$ is slightly abusive as it implies weak* convergence when what is meant is the *weak convergence of measures*. Yet it should be emphasized that the two notions are *equivalent* on the space of *probability measures*.

The above definitions and theorem provide several characterizations of the weak* topology on the set of probability measures. While this characterization is mathematically sufficient, our discussions of approximation and convergence would be facilitated by some metric ρ defined on the space $\mathcal{P}(\Theta)$ which metrizes the above notions of topological convergence. That is, given two probability measures P and Q, we would like ρ to have the property that $Q \in B_\epsilon(P)$ if and only if $\rho(P,Q) < \epsilon$. Such a metric could then be used to establish more intuitive notions of convergence, compactness, etc., in the space of probability measures. In fact, such a metric does exist, named for the Russian probabilist Y.V. Prohorov who first defined the metric [29] and derived its properties.

Definition 2.5. *Let (Θ, d) be a metric space. For all $F \in \Sigma_\Theta$, $F \neq \emptyset$, define the ϵ-neighborhood of F,*

$$F^\epsilon = \{\phi \in \Theta | \inf_{\theta \in \Theta} d(\theta, \phi) < \epsilon\}.$$

If $F = \emptyset$, define $F^\epsilon = \emptyset$.

Definition 2.6. *Let (Θ, d) be a metric space and let $\mathcal{P}(\Theta)$ be the set of all probability measures on Θ. For any two measures $P, Q \in \mathcal{P}(\Theta)$, the Prohorov metric ρ is*

$$\rho(P,Q) = \inf\left\{\epsilon > 0 | Q(F) \leq P(F^\epsilon) + \epsilon \text{ and } P(F) \leq Q(F^\epsilon) + \epsilon, \text{ for all } F \in \Sigma_\Theta\right\}.$$

This definition of the Prohorov metric is far from intuitive. We will first prove that Definition 2.6 does indeed describe a valid metric. Next, we show that ρ metrizes the weak* topology.

Theorem 2.7. *Let (Θ, d) be a separable metric space. Then ρ is a metric on $\mathcal{P}(\Theta)$.*

Proof. By construction, ρ is nonnegative and symmetric and $\rho(P,Q) = 0$ if $P = Q$. We must show $\rho(P,Q) = 0$ implies $P = Q$, and that ρ is subadditive.

Assume $\rho(P,Q) = 0$. Then $P(F) = Q(F)$ for all $F \in \Sigma_\Theta$ (and, in particular, for all closed sets in Θ). Since (Θ, d) is separable, all probability measures on Θ are regular [25, pg. 7], and thus are uniquely determined by their values on closed sets. Thus we may conclude $P = Q$.

To show subadditivity, assume $\rho(P_1, P_2) = \epsilon_1$ and $\rho(P_2, P_3) = \epsilon_2$. We need to show $\rho(P_1, P_3) \leq \epsilon_1 + \epsilon_2$. From the definition of ρ, the following inequalities hold for all $F \in \Sigma_\Theta$:

$$\begin{aligned} P_1(F) &\leq P_2(F^{\epsilon_1}) + \epsilon_1 \\ P_2(F) &\leq P_1(F^{\epsilon_1}) + \epsilon_1 \\ P_2(F) &\leq P_3(F^{\epsilon_2}) + \epsilon_2 \\ P_3(F) &\leq P_2(F^{\epsilon_2}) + \epsilon_2. \end{aligned}$$

Thus we have

$$\begin{aligned} P_1(F) &\leq P_2(F^{\epsilon_1}) + \epsilon_1 \leq P_3\left((F^{\epsilon_1})^{\epsilon_2}\right) + \epsilon_1 + \epsilon_2 \\ P_3(F) &\leq P_2(F^{\epsilon_2}) + \epsilon_2 \leq P_1\left((F^{\epsilon_2})^{\epsilon_1}\right) + \epsilon_2 + \epsilon_1 \end{aligned}$$

Trivally, $(F^{\epsilon_1})^{\epsilon_2} \subset F^{\epsilon_1+\epsilon_2}$. Hence $P_1(F) \le P_3(F^{\epsilon_1+\epsilon_2}) + \epsilon_1 + \epsilon_2$ and $P_3(F) \le P_1(F^{\epsilon_1+\epsilon_2}) + \epsilon_1 + \epsilon_2$. Since these statements hold for all $F \in \Sigma_\Theta$, $\rho(P_1, P_3) \le \epsilon_1 + \epsilon_2$. □

Theorem 2.8. *Assume (Θ, d) is separable. Assume $P_M \in \mathcal{P}(\Theta)$ for all $M = 1, \ldots, \infty$, and $P \in \mathcal{P}(\Theta)$. Then $P_M \xrightarrow{w^*} P$ if and only if $\rho(P_M, P) \to 0$.*

Proof. ($\Leftarrow$) Assume $\rho(P_M, P) \to 0$. Then for all $\epsilon > 0$ there exists $\tilde{M} = \tilde{M}(\epsilon)$ such that

$$P_M(F) < P(F^\epsilon) + \epsilon \quad \text{and} \quad P(F) < P_M(F^\epsilon) + \epsilon$$

for all $F \in \Sigma_\Theta$. Let C by any closed set in Θ. (Then $C \in \Sigma_\Theta$.) Since (Θ, d) is separable, P is regular and there exists $\delta < \epsilon$ such that

$$P(C^\delta) < P(C) + \frac{\epsilon}{2}.$$

Take $\tilde{M} = \tilde{M}(\delta/2)$, Then $\rho(P_M, P) < \frac{\delta}{2}$ for all $M > \tilde{M}$ and

$$\begin{aligned} P_M(C) &< P(C^{\delta/2}) + \frac{\delta}{2} && \text{(defn of } \rho) \\ &< P(C) + \frac{\epsilon}{2} + \frac{\delta}{2} && \text{(regularity of } P) \\ &< P(C) + \epsilon && \text{(choice of } \delta). \end{aligned}$$

Hence $\limsup P_M(C) \le P(C)$ for all C closed in Θ and $P_M \xrightarrow{w^*} P$ by Definition 2.4.

($\Rightarrow$) Assume $P_M \xrightarrow{w^*} P$. For all $\epsilon > 0$, fix δ such that $0 < \delta < \frac{\epsilon}{3}$. By the separability of Θ, there exist open sets $B_\delta(\theta_k)$ such that

$$\bigcup_{k=1}^{\infty} B_\delta(\theta_k) = \Theta.$$

Fix n_0 such that

$$P\left(\bigcup_{k=1}^{n_0} B_\delta(\theta_k)\right) \ge 1 - \delta. \tag{2.2}$$

(Such a a value n_0 must exist since $\lim_{n\to\infty} P\left(\bigcup_{k=1}^{n} B_\delta(\theta_k)\right) = 1$.) Define the collection of all possible (nonempty) unions of the sets $B_\delta(\theta_k)$, $1 \le k \le n_0$,

$$\mathcal{O} = \left\{ \bigcup_K B_\delta(\theta_k) \middle| K \subset \{1, 2, \ldots, n_0\} \right\}.$$

Then for all $A \in \mathcal{O}$, $A \in \Sigma_\Theta$ and $\partial A \subset \bigcup_{k=1}^{n_0} \partial B_\delta(\theta_k)$ so that $P(\partial A) = 0$. Then A is a P-continuity set and $P_M(A) \to P(A)$ by assumption. Thus there exists $\tilde{M}$ such that

$$|P_M(A) - P(A)| < \delta \tag{2.3}$$

for all $M > \tilde{M}$ and for all $A \in \mathcal{O}$. In particular, for $A = \bigcup_{k=1}^{n_0} B_\delta(\theta_k)$,

$$\begin{aligned} P_M\left(\bigcup_{k=1}^{n_0} B_\delta(\theta_k)\right) &\geq P\left(\bigcup_{k=1}^{n_0} B_\delta(\theta_k)\right) - \delta \\ &\geq 1 - 2\delta \qquad \text{(by (2.3))}. \end{aligned} \tag{2.4}$$

Now, we need to show $P_M(F) < P(F^\epsilon) + \epsilon$ and $P(F) < P_M(F^\epsilon) + \epsilon$ for all $F \in \Sigma_\Theta$. Let $F \in \Sigma_\Theta$ be arbitrary. Define

$$A = \bigcup \left\{ B_\delta(\theta_k) \middle| B_\delta(\theta_k) \bigcap F \neq \emptyset \right\}$$

where the union is taken over $1 \leq k \leq n_0$. The following facts are trivially verified:

$$A \in \mathcal{O} \tag{2.5}$$

$$A \subset F^\delta \tag{2.6}$$

$$F \subset A \bigcup \left(\bigcup_{k=1}^{n_0} B_\delta(\theta_k)\right)^C. \tag{2.7}$$

Then for all $M \geq \tilde{M}$,

$$\begin{aligned} P(F) &\leq P(A) + P\left(\left(\bigcup_{k=1}^{n_0} B_\delta(\theta_k)\right)^C\right) && \text{(by (2.7))} \\ &\leq P(A) + \delta && \text{(by (2.2))} \\ &\leq P_M(A) + 2\delta && \text{(by (2.3))} \\ &\leq P_M(F^\delta) + 2\delta && \text{(by (2.5))} \\ &\leq P_M(F^\epsilon) + \epsilon && \text{(by choice of } \delta). \end{aligned}$$

and

$$\begin{aligned} P_M(F) &\leq P_M(A) + P_M\left(\left(\bigcup_{k=1}^{n_0} B_\delta(\theta_k)\right)^C\right) && \text{(by (2.7))} \\ &\leq P_M(A) + 2\delta && \text{(by (2.4))} \\ &\leq P(A) + 3\delta && \text{(by (2.3))} \\ &\leq P(F^\delta) + 3\delta && \text{(by (2.5))} \\ &\leq P(F^\epsilon) + \epsilon && \text{(by choice of } \delta). \end{aligned}$$

Hence $\rho(P_M, P) \leq \epsilon$ for all $M \geq \tilde{M}$. □

With these considerations, we have obtained the desired result–the weak topology of measures (weak* topology) is equivalent to the topology induced by the Prohorov metric on the space of probability measures over a separable metric space (Θ, d). It should be noted that in the definition of the Prohorov metric, it is sufficient to consider only sets F which are closed (see [33, Online Supplement] for a proof; this follows from the fact that probability measures are regular [25, pg. 7]), so that the definitions and results presented here are in agreement with similar results obtained previously [4, 6, 11, 16, 21]. We now proceed to use the Prohorov metric to establish a list of propositions, theorems and corollaries which will be of use as we return to the original problem of setting up a least-squares estimation framework for the nonparametric estimation of measures.

3 Some Useful Results

From the results of the previous section, we know that given a separable metric space (Θ, d), the space $(\mathcal{P}(\Theta), \rho)$ of probability measures on Θ is a metric space with a topology equivalent to the weak topology of measures (weak* topology). We now focus on characterising the properties of the space $(\mathcal{P}(\Theta), \rho)$ which will prove useful in establishing results for the parameter estimation problem.

Define

$$D = \{\Delta_{\theta_k} | \theta_k \in \Theta\}.$$

That is, D is the space of Dirac measures on Θ, defined for all $F \in \Sigma_\Theta$ as

$$\Delta_{\theta_k}(F) = \begin{cases} 1, & \theta_k \in F \\ 0, & \theta_k \notin F \end{cases}$$

Proposition 3.1. *Let (Θ, d) be a separable metric space and define $D \subset \mathcal{P}(\Theta)$ as above. Then*

$$\rho(\Delta_{\theta_1}, \Delta_{\theta_2}) = \min\{d(\theta_1, \theta_2), 1\}.$$

Proof. Note first that $\rho(P, Q) \leq 1$ for all $P, Q \in \mathcal{P}(\Theta)$. Take $\epsilon > d(\theta_1, \theta_2)$. Then for all $F \in \Sigma_\Theta$, $\theta_1 \in F \Rightarrow \theta_2 \in F^\epsilon$ and $\theta_2 \in F \Rightarrow \theta_1 \in F^\epsilon$. Thus for all $F \in \Sigma_\Theta$,

$$\begin{aligned} \Delta_{\theta_1}(F) &\leq \Delta_{\theta_2}(F^\epsilon) + \epsilon \\ \Delta_{\theta_2}(F) &\leq \Delta_{\theta_1}(F^\epsilon) + \epsilon. \end{aligned}$$

Thus $\rho(\Delta_{\theta_1}, \Delta_{\theta_2}) \leq \epsilon$. Since this holds for all $\epsilon > d(\theta_1, \theta_2)$, we have $\rho(\Delta_{\theta_1}, \Delta_{\theta_2}) \leq d(\theta_1, \theta_2)$.

Now take ϵ such that $\rho(\Delta_{\theta_1}, \Delta_{\theta_2}) < \epsilon < 1$. Then

$$\begin{aligned} \Delta_{\theta_1}(F) &< \Delta_{\theta_2}(F^\epsilon) + \epsilon \\ \Delta_{\theta_2}(F) &< \Delta_{\theta_1}(F^\epsilon) + \epsilon \end{aligned}$$

for all $F \in \Sigma_\Theta$. Take $F = \{\theta_1\}$. Then the first inequality above implies $1 < \Delta_{\theta_2}(B_\epsilon(\theta_1)) + \epsilon$. Since $\epsilon < 1$, we must have $\Delta_{\Theta_1}(B_\epsilon(\theta_1)) > 0$ and thus $\theta_2 \in B_\epsilon(\theta_1)$. Hence $d(\theta_1, \theta_2) < \epsilon$. Since this holds for all $\rho(\Delta_{\theta_1}, \Delta_{\theta_2}) < \epsilon < 1$, we must have $\min\{d(\theta_1, \theta_2), 1\} \leq \rho(\Delta_{\theta_1}, \Delta_{\theta_2})$. Hence the stated result holds. □

Corollary 3.2. *The sequence $\{\theta_k\}_{k=1}^\infty$ is Cauchy in the separable space (Θ, d) if and only if the sequence $\{\Delta_{\theta_k}\}_{k=1}^\infty$ is Cauchy in $(\mathcal{P}(\Theta, \rho)$.*

Proof. Trivial, by previous proposition. □

Corollary 3.3. *Let (Θ, d) be a separable metric space and let the space D be defined as above. Then D is sequentially closed in $(\mathcal{P}(\Theta), \rho)$. (That is, D is weak* sequentially closed in the space of probability measures.)*

Proof. Assume the sequence $\{\Delta_{\theta_k}\}_{k=1}^{\infty}$ converges in the Prohorov metric to some $P \in \mathcal{P}(\Theta)$. We need to show $P \in D$. An obvious candidate is $P = \Delta_\theta$ where $\theta = \lim \theta_k$, if such a limit were to exist. We show that this is the case.

Consider the sequence $\{\theta_k\}_{k=1}^{\infty}$ and assume (for the purpose of reaching a contradiction) that this sequence does not have a convergent subsequence. (Then any element in the sequence $\{\theta_k\}_{k=1}^{\infty}$ can be repeated at most only a finite number of times, and we may assume without loss of generality that no element of the sequence is repeated.) Define the set $S = \{\theta_1, \theta_2, \ldots\}$. Then S is (vacuously) closed in Θ, as is every subset of S. Now consider any subsequence $C = \{\theta_{k_1}, \theta_{k_2}, \ldots\} \subset S$. Then by the weak convergence of the set $\{\Delta_{\theta_k}\}_{k=1}^{\infty}$, and because C cannot contain a convergent subsequence

$$P(C) \geq \limsup \Delta_{\theta_k}(C) = 1.$$

However, define C_k to be the subset obtained by removing the element θ_k from S. (If θ_k were repeated n_k times, one obtains the same result by removing all instances of θ_k.) Then $P(C_k) = 1$ for all k (by the argument above), and hence $P(\{\theta_k\}) = 0$ for all k. But S is the disjoint countable union of the point sets $\{\theta_k\}$, hence we would have $P(S) = 0$. But $P(S) = 1$ since S is itself a closed set. Thus we have reached a contradiction.

So the sequence $\{\theta_k\}_{k=1}^{\infty}$ must have a convergent subsequence, $\theta_{k_l} \to \theta$. But then we must have $\Delta_{\theta_k} \to \Delta_\theta$ by Corollary 3.2. Hence $P = \Delta_\theta$ by the uniqueness of weak* limits. □

Definition 3.4. *$P \in \mathcal{P}(\Theta)$ is tight if for all $\epsilon > 0$ there exists a compact set $K \subset \Theta$ such that $P(K) > 1 - \epsilon$. A family of measures $\Pi \subset \mathcal{P}(\Theta)$ is tight if for all $P \in \Pi$, P is tight.*

Theorem 3.5. *Assume (Θ, d) is complete. If for all $\epsilon, \delta > 0$ there exist $\theta_1, \ldots, \theta_M \in \Theta$ such that*

$$P\left(\bigcup_{k=1}^{M} B_\delta(\theta_k)\right) \geq 1 - \epsilon,$$

for all $P \in \Pi$, then Π is tight.

Proof. For all $\epsilon > 0$, for each $n \geq 1$, take $\delta = 1/n$. By hypothesis, there exist $\theta_1^n, \ldots, \theta_{M_n}^n \in \Theta$ such that

$$P\left(\bigcup_{k=1}^{M_n} B_{1/n}(\theta_k^n)\right) \geq 1 - 2^{-n}\epsilon$$

for all $P \in \Pi$. Define

$$K = \bigcap_{n=1}^{\infty} \bigcup_{k=1}^{M_n} \bar{B}_{1/n}(\theta_k^n).$$

Then K is closed and for $\tilde{n} > 1/\delta$,

$$K \subset \bigcup_{k=1}^{M_{\tilde{n}}} \bar{B}_{1/\tilde{n}}(\theta_k^{\tilde{n}}) \subset \bigcup_{k=1}^{M_{\tilde{n}}} B_\delta(\theta_k^{\tilde{n}}).$$

Thus K is a totally bounded subset of a complete space. Thus K is compact. Moreover, for any $P \in \Pi$,

$$\begin{aligned}
P(K) &= \lim_{N\to\infty} P\left(\bigcap_{n=1}^{N} \bigcup_{k=1}^{M_n} \bar{B}_{1/n}(\theta_k^n)\right) \\
&= 1 - \lim_{N\to\infty} P\left(\bigcup_{n=1}^{N} \left[\bigcup_{k=1}^{M_n} \bar{B}_{1/n}(\theta_k^n)\right]^C\right) \\
&\geq 1 - \lim_{N\to\infty} \sum_{n=1}^{N} P\left(\left[\bigcup_{k=1}^{M_n} \bar{B}_{1/n}(\theta_k^n)\right]^C\right) \\
& \quad 1 - \sum_{n=1}^{\infty} 2^{-n}\epsilon \\
&= 1 - \epsilon.
\end{aligned}$$

□

Theorem 3.6 (Prohorov)**.** *Assume (Θ, d) is separable and let $\Pi \subset (\mathcal{P}(\Theta), \rho)$. The following are equivalent:*

- Π *is relatively (sequentially) compact;*
- Π *is tight.*

Proof. See [25, Ch. 1.6] □

Corollary 3.7. *Assume (Θ, d) is separable. Then (Θ, d) is complete if and only if $(\mathcal{P}(\Theta), \rho)$ is complete.*

Proof. ($\Rightarrow$) Assume $\{P_M\}_{M=1}^{\infty}$ is a weak* Cauchy sequence in $(\mathcal{P}(\theta), \rho)$. We need to show there exists some $P \in \mathcal{P}(\Theta)$ such that $P_M \xrightarrow{w^*} P$. To do so, it is sufficient to show that $\{P_M\}_{M=1}^{\infty}$ has at least one convergent subsequence. If we can show that the collection of measures is tight, then it is weak* relatively sequentially compact in $\mathcal{P}(\Theta)$ by Prohorov's Theorem (Theorem 3.6). To prove the tightness of this collection of measures, we will use Theorem 3.5.

Let $\{\theta_1, \theta_2, \ldots\}$ be an enumeration of the countable, dense subset of Θ. For all $\epsilon, \delta > 0$, fix $\eta < \min\{\epsilon, \delta\}/2$. Since the sequence $\{P_M\}_{M=1}^{\infty}$ is a weak* Cauchy, there exists $\tilde{M} = \tilde{M}(\eta)$ such that $P_M(F) \leq P_N(F^\eta) + \eta$ and $P_N(F) \leq P_M(F^\eta) + \eta$ for all $M, N > \tilde{M}$ and for all $F \in \Sigma_\Theta$.

Note that, by construction,

$$\bigcup_{k=1}^{\infty} B_{\delta/2}(\theta_k) = \Theta.$$

Hence for each $1 \leq M \leq \tilde{M}$,

$$\lim_{n \to \infty} P_M \left(\bigcup_{k=1}^{n} B_{\delta/2}(\theta_k) \right) = 1$$

and there exists an n_0 such that

$$P_M \left(\bigcup_{k=1}^{n_0} B_{\delta/2}(\theta_k) \right) \geq 1 - \eta. \tag{3.1}$$

(Such an n_0 must exist separately for each value of M, of which there are a finite number.) Now note that

$$\left(\bigcup_{k=1}^{n_0} B_{\delta/2}(\theta_k) \right)^{\eta} \subset \bigcup_{k=1}^{n_0} B_{\delta/2+\eta}(\theta_k) \subset \bigcup_{k=1}^{n_0} B_{\delta}(\theta_k). \tag{3.2}$$

Hence for all $M \geq \tilde{M}$,

$$\begin{aligned} P_{\tilde{M}} \left(\bigcup_{k=1}^{n_0} B_{\delta/2}(\theta_k) \right) &\leq P_M \left(\left(\bigcup_{k=1}^{n_0} B_{\delta/2}(\theta_k) \right)^{\eta} \right) + \eta && \text{(by defn of } \tilde{M}\text{)} \\ &\leq P_M \left(\bigcup_{k=1}^{n_0} B_{\delta}(\theta_k) \right) + \eta && \text{(by (3.2))} \end{aligned}$$

and therefore

$$\begin{aligned} P_M \left(\bigcup_{k=1}^{n_0} B_{\delta}(\theta_k) \right) &\geq P_{\tilde{M}} \left(\bigcup_{k=1}^{n_0} B_{\delta/2}(\theta_k) \right) - \eta \\ &\geq 1 - 2\eta && \text{(by (3.1))} \\ &\geq 1 - \epsilon && \text{(by choice of } \eta\text{).} \end{aligned}$$

Finally, for $1 \leq M \leq \tilde{M}$

$$\begin{aligned} P_M \left(\bigcup_{k=1}^{n_0} B_{\delta}(\theta_k) \right) &\geq P_M \left(\bigcup_{k=1}^{n_0} B_{\delta/2}(\theta_k) \right) \\ &\geq 1 - \eta && \text{(by (3.1))} \\ &\geq 1 - \epsilon && \text{(by choice of } \eta\text{).} \end{aligned}$$

So P_M is tight for all $M = 1, \ldots, \infty$. Hence $\{P_M\}_{M=1}^{\infty}$ is tight and thus relatively compact in $\mathcal{P}(\Theta)$ and so has a convergent subsequence to some $P \in \mathcal{P}(\Theta)$.

($\Leftarrow$) Assume $\{\theta_k\}_{k=1}^{\infty}$ is Cauchy in (Θ, d). We need to show $\theta_k \to \theta$ for some $\theta \in \Theta$. Since $\{\theta_k\}_{k=1}^{\infty}$ is Cauchy, the sequence of Dirac measures $\{\Delta_{\theta_k}\}_{k=1}^{\infty}$ is also Cauchy by Corollary 3.2. By the completeness of $(\mathcal{P}(\Theta), \rho)$, there exists $P \in \mathcal{P}(\Theta)$ such that $\Delta_{\theta_k} \xrightarrow{w^*} P$. But D (the space of all Dirac measures) is closed by Corollary 3.3. Hence $P = \Delta_\theta$, for some $\theta \in \Theta$. Hence $\theta_k \to \theta$. □

Corollary 3.8. *Assume (Θ, d) is separable. Then (Θ, d) is compact if and only if $(\mathcal{P}(\Theta), \rho)$ is compact.*

Proof. ($\Rightarrow$) If (Θ, d) is compact then every collection of measures on Θ (and specifically $\mathcal{P}(\Theta)$ itself) is tight and thus relatively compact by Theorem 3.6. Since $(\Theta$,d) is compact, it is also complete and so is $(\mathcal{P}(\Theta), \rho)$ (by the previous corollary) so that $(\mathcal{P}(\Theta, \rho)$ must be closed. Hence relative compactness is compactness.

($\Leftarrow$) See the proof of the converse half of the previous corollary; given an arbitrary sequence $\{\theta_k\}_{k=1}^{\infty}$, it must have a convergent subsequence. □

It is interesting to note that we may revisit the Riesz Representation Theorem (Theorem 2.1) for an alternative proof of the direct half of the previous corollary. Given the compactness of (Θ, d), by the Riesz Representation Theorem we have $P_M \xrightarrow{w^*} P$ if and only if $f^*_{P_M}(f) \to f^*_P(f)$ for all $f \in C_B(\Theta)$. Now, consider the ball

$$B = \left\{ f^* \in C_B(\Theta)^* \Big| \, ||f^*|| \leq 1 \right\}.$$

This is the unit ball in $C_B(\Theta)^*$, which is compact in the weak* topology by Alaoglu's Theorem [30, pg. 237]. We may then observe that

$$\left\{ f^* \in B \Big| \, ||f^*|| = 1, \text{ and } f^* \text{ positive} \right\}$$

is homeomorphic to $(\mathcal{P}(\Theta, \rho)$. This set is also closed in B, and hence compact.

The compactness of the space $(\mathcal{P}(\Theta), \rho)$ given the compactness of (Θ, d) is of vital importance for the theoretical framework to be discussed in the next sections. In effect, one need only show that the cost functional $J_n(\vec{v}, P)$ in (1.6) is a continuous function of P in order to be guaranteed the existence of a minimizer to the least squares estimation problem.

We need one final result which will be useful in establishing computational tools for the parameter estimation problem.

Theorem 3.9. *Assume (Θ, d) is a separable, compact metric space. Let $\Theta_d = \{\theta_k\}_{k=1}^{\infty}$ be an enumeration of the countable dense subset of Θ. Take $\mathbb{Q} \subset \mathbb{R}$ to be the set of all rational numbers. Define*

$$\tilde{\mathcal{P}}_d(\Theta) = \left\{ P \in \mathcal{P}(\Theta) \Big| P = \sum_{k=1}^{M} p_k \Delta_{\theta_k}, \theta_k \in \Theta_d, M \in \mathbb{N}, p_k \in [0,1] \cap \mathbb{Q}, \sum_{k=1}^{M} p_k = 1 \right\}.$$

(That is, $\tilde{\mathcal{P}}_d(\Theta)$ is the collection of all convex combinations of Dirac measures on Θ with atoms $\theta_k \in \Theta_d$ and rational weights.) Then $\tilde{\mathcal{P}}_d(\Theta)$ is dense in $\mathcal{P}(\Theta)$, and thus $\mathcal{P}(\Theta)$ is separable.

Proof. $\tilde{\mathcal{P}}_d(\Theta)$ is obviously countable. Let $\epsilon > 0$ and let $P \in \mathcal{P}(\Theta)$ be arbitrary. We need to show there exists $P_M \in \tilde{\mathcal{P}}_d(\Theta)$ such that $\rho(P_M, P) < \epsilon$. As before, we first note that for each $M \geq 1$,

$$\bigcup_{k=1}^{\infty} B_{1/M}(\theta_k) = \Theta$$

so that we may choose $n_0 = n_0(M)$ satisfying

$$P\left(\bigcup_{k=1}^{n_0} B_{1/M}(\theta_k)\right) \geq 1 - 1/M.$$

Define

$$\begin{aligned} A_1^M &= B_{1/M}(\theta_1) \\ A_k^M &= B_{1/M}(\theta_k) - \bigcup_{j=1}^{k-1} B_{1/M}(\theta_j), \quad k = 2, \ldots n_0. \end{aligned}$$

Then the sets A_k^M, $1 \leq k \leq n_0$ are disjoint and

$$\begin{aligned} \bigcup_{k=1}^{n} A_k^M &= \bigcup_{k=1}^{n} B_{1/M}(\theta_k), \quad 1 \leq n \leq n_0 \\ P\left(\bigcup_{k=1}^{n_0} A_k^M\right) &\geq 1 - \frac{1}{M}. \end{aligned}$$

Pick the values $p_k^n \in [0,1] \cap \mathbb{Q}$ such that

$$\begin{aligned} \sum_{k=1}^{n_0} p_k^M &= 1 \\ \sum_{k=1}^{n_0} \left|P(A_k^M) - p_k^M\right| &< \frac{2}{M}. \end{aligned}$$

(To do so, one may first freely choose values $\hat{p}_k^M \in [0,1] \cap \mathbb{Q}$ such that

$$\sum_{k=1}^{n_0} \left|P(A_k^M) - \hat{p}_k^M\right| < \frac{1}{2M}$$

and then set $p_k^M = \hat{p}_k^M / \sum_{k=1}^{n_0} \hat{p}_k^M$.) Now define

$$P_M = \sum_{k=1}^{n_0} p_k^M \Delta_{\theta_k}.$$

We must show that $\rho(P_M, P) \to 0$ as M gets large. For any $f \in C_B(\Theta)$,

$$\begin{aligned}
\left|\int_\Theta f(\theta)dP_M(\theta) - \int_\Theta f(\theta)dP(\theta)\right| &= \left|\sum_{k=1}^{n_0} p_k^M f(\theta_k) - \int_\Theta f(\theta)dP(\theta)\right| \\
&\leq \left|\sum_{k=1}^{n_0} P(A_k^M) f(\theta_k) - \int_\Theta f(\theta)dP(\theta)\right| + \frac{2}{M}\sup_k |f(\theta_k)| \\
&\leq \left|\int_\Theta \sum_{k=1}^{n_0} f(\theta_k)\chi(\theta)_{A_k^M} dP(\theta) - \int_\Theta f(\theta)dP(\theta)\right| + \frac{2}{M}\,||f||_\infty \\
&\leq \left|\sum_{k=1}^{n_0} \int_\Theta \left(f(\theta_k)\chi(\theta)_{A_k^M} - f(\theta)\chi(\theta)_{A_k^M}\right) dP(\theta)\right. \\
&\qquad \left. - \int_\Theta f(\theta)\chi_{(\cup_{k=1}^{n_0} A_k^M)^C} dP(\theta)\right| + \frac{2}{M}\,||f||_\infty \\
&\leq \sum_{k=1}^{n_0} \sup_{\theta\in A_k^M} |f(\theta_k) - f(\theta)|\, P(A_k^M) \\
&\qquad + ||f||_\infty\, P\left(\left(\bigcup_{k=1}^{n_0} A_k^M\right)^C\right) + \frac{2}{M}\,||f||_\infty\,.
\end{aligned}$$

(The function $\chi(\theta)_A$ is the indicator function on the set A.) Recall $A_k^M \subset B_{1/M}(\theta_k)$ by construction. Thus $\theta \in A_k^M$ implies $d(\theta_k, \theta) < 1/M$ and for M large enough, $|f(\theta_k) - f(\theta)| < \epsilon$ for all $\theta \in A_k^M$ and for all k (since $f \in C_B(\Theta)$ for Θ compact and thus f is uniformly continuous). Altogether we have

$$\left|\int_\Theta f(\theta)dP_M(\theta) - \int_\Theta f(\theta)dP(\theta)\right| \leq \epsilon + \frac{||f||_\infty}{M} + \frac{2\,||f||_\infty}{M}$$

and the result is proved. □

An alternative proof of the above result can be found in [4].

4 Existence and Consistency of the Estimator

We now turn our attention to characterizing the least squares estimator (1.5) and its corresponding estimate (1.6). In the present section we ignore any computational approximations and establish results concerning the theoretical existence and consistency of the least squares estimator and estimate, regardless of our ability to compute them (although the method of proof does foreshadow the computational approach in the next section).

4.1 Existence of the Estimator

We begin by proving the existence of P_n and $\hat{P}_n$ as measurable functions mapping a subset of $\mathbb{R}^n$ (that is, the data) into the space of probability measures on Θ. We remark that the statement of Theorem 4.1 concerns the estimate $\hat{P}_n$ obtained from the data realizations $\vec{v} \in \mathbb{R}^n$. This is sufficient to establish the existence of the estimator P_n as a measurable function as well, since the random vector $\vec{V}$ is by definition a measurable function from a probability triple into $\mathbb{R}^n$, and the composition of measurable functions is measurable.

Theorem 4.1. *Define the function $J_n : \mathbb{R}^n \times \mathcal{P}(\Theta) \to \mathbb{R}$ according to Equation* (1.6). *Assume (Θ, d) is separable and compact and take the space of probability measures $\mathcal{P}(\Theta)$ with the Prohorov metric ρ. Assume further that $J_n(\cdot, P)$ is a measurable function from $\mathbb{R}^n \to \mathbb{R}$ for each $P \in \mathcal{P}(\Theta)$, and that $J_n(\vec{v}, \cdot) : \mathcal{P}(\Theta) \to \mathbb{R}$ is continuous for each $\vec{v} \in \mathbb{R}^n$. Then there exists a measurable function $\hat{P}_n : \mathbb{R}^n \to \mathcal{P}(\Theta)$ such that*

$$J(\vec{v}, \hat{P}_n(\vec{v})) = \inf_{P \in \mathcal{P}(\Theta)} J(\vec{v}, P).$$

Proof. Let $\Theta_d = \{\theta_k\}_{k=1}^{\infty}$ be an enumeration of the countable dense subset of Θ as used in Theorem 3.9. For each $M \geq 1$, define

$$\mathcal{P}_M(\Theta) = \left\{ P \in \tilde{\mathcal{P}}_d(\Theta) \middle| P = \sum_{k=1}^{M} p_k \Delta_{\theta_k}, \theta_k \in \{\theta_i\}_{i=1}^{M} \right\} \subset \tilde{\mathcal{P}}_d(\Theta). \tag{4.1}$$

(That is, $\mathcal{P}_M$ is the set of all discrete measures consisting of a convex combination of M Dirac measures with atoms in $\{\theta_i\}_{i=1}^{M}$ weighted with rational coefficients.) Thus $\mathcal{P}_M$ is countable. Let $\{P_j^M\}_{j=1}^{\infty}$ be an enumeration of the elements of $\mathcal{P}_M$. (We remark that, because the M nodes θ_k are fixed in advance, the space $\mathcal{P}_M$ can be analogously considered as a subset of $\mathbb{R}^M$, a fact which will be exploited in some of the notation below.) Finally, define $\mathcal{P}_J^M = \{P_j^M\}_{j=1}^{J}$, the first J enumerated elements of $\mathcal{P}_M$.

Fix $J \geq 1$. Define the function $\tilde{P}_J^M(\vec{v})$ implicitly as

$$J(\vec{v}, \tilde{P}_J^M(\vec{v})) = \min_{P \in \mathcal{P}_J^M} J(\vec{v}, P).$$

Such a function must exist because the minimum is begin taken over a finite number of elements from a point set; if the minimum occurs at multiple elements of $\mathcal{P}_J^M$, we may arbitrarily choose the element which comes first in the enumeration so that the function $\tilde{P}_J^M(\vec{v})$ is well-defined. First, we show that $\tilde{P}_J^M(\vec{v})$ is measurable.

Let $F \in \Sigma_{P_J^M}$. (Thus F is a finite point set.) We must show that the set B defined as

$$B = \left\{ \vec{v} \middle| \tilde{P}_J^M(\vec{v}) \in F \right\}$$

is contained within measurable sets $\Sigma_{\mathbb{R}^n}$ in $\mathbb{R}^n$. Since F is a finite point set, we can define for each $P_j^M \in F$ the sets

$$\begin{aligned} B_j &= \left\{ \vec{v} \middle| \tilde{P}_J^M(\vec{v}) = P_j^M \right\} \\ &= \left\{ \vec{v} \middle| J(\vec{v}, P_j^M(\vec{v})) = \min_{P \in \mathcal{P}_J^M} J(\vec{v}, P) \right\} \\ &= \left\{ \vec{v} \middle| J(\vec{v}, P_j^M(\vec{v})) = \min_{1 \le j \le J} J(\vec{v}, P_j^M) \right\}. \end{aligned}$$

By assumption, the functions $J(\vec{v}, P_j^M)$ are measurable from $\mathbb{R}^n$ into $\mathbb{R}$ for all P_j^M, $j \geq 1$. The minimum over a finite set of functions is also measurable, as is the test for equality. Hence $B_j \in \Sigma_{\mathbb{R}^n}$. Finally, $B = \cup B_j$, the union being over the finite number of sets B_j, hence $B \in \Sigma_{\mathbb{R}^n}$ and the function $\tilde{P}_J^M(\vec{v})$ is measurable.

As mentioned previously, we can identify the function $\tilde{P}_J^M(\vec{v})$ with $[0,1]^M \cap \mathbb{Q}^M$ via the map $\tilde{P}_J^M(\vec{v}) \mapsto (p_1^M(\vec{v}), \ldots, p_M^M(\vec{v}))$. Let $\tilde{p}_J^M$ be the first component of the vector representation for $\tilde{P}_J^M(\vec{v})$. Now consider the sequence $\{\tilde{p}_J^M\}_{J=1}^\infty$. Define

$$\hat{p}_1^M(\vec{v}) = \liminf_{J \to \infty} \tilde{p}_J^M(\vec{v}).$$

Since each $\tilde{p}_J^M(\vec{v})$ is a measurable function, so is $\hat{p}_1^M(\vec{v})$. Also, since the space $[0,1]^M$ is compact, there must exist a convergent subsequence $\tilde{P}_{J_l}^M$ of (the vector representation of) $\tilde{P}_J^M$ to some vector $(\hat{p}_1^M(\vec{v}), \bar{p}_2^M(\vec{v}), \ldots, \bar{p}_M^M(\vec{v}))$, which can be identified with a measure $\bar{P}_M$. Now

$$\begin{aligned} \inf_{[0,1]^{M-1} \cap \mathbb{Q}^{M-1}} J_n(\vec{v}, (\hat{p}_1^M, p_2, \ldots, p_M)) &\le J_n(\vec{v}, \bar{P}_M) \\ &= \lim_l J_n(\vec{v}, \tilde{P}_{J_l}^M) \\ &= \lim_l \inf_{P \in \mathcal{P}_{J_l}^M} J_n(\vec{v}, P) \\ &= \inf_{P \in \mathcal{P}_M} J_n(\vec{v}, P). \end{aligned}$$

The first equality comes from the definition of $\tilde{P}_M$ and the continuity of the function J; the second equality comes from the definition of $\bar{P}_M$ as the limit of the probability measures $\tilde{P}_{J_l}^M$; the final equality arises from the density of $\{P_j^M\}$ in $\mathcal{P}$.

Now, define (with some abuse of notation)

$$J_n^{(1,M)}(\vec{v}, P) = J_n(\vec{v}, (\hat{p}_1^M, p_2, \ldots, p_M)).$$

Applying the same arguments above inductively on $J_n^{(j,M)}$, we obtain a set of measurable functions $\hat{p}_1^M(\vec{v}), \ldots, \hat{p}_M^M(\vec{v})$ such that

$$J_n(\vec{v}, (\vec{p}_1^M, \ldots, \hat{p}_M^M)) = \inf_{P \in \mathcal{P}_M} J_n(\vec{v}, P)$$

and we have proven the existence of a measurable function $\hat{P}_M \in \mathcal{P}_M$ mapping $\mathbb{R}^n \to \mathcal{P}(\Theta)$ which minimizes the cost functional J_n. We conclude the proof by noting that

$$J_n(\vec{v}, \hat{P}(\vec{v})) = \inf_{P \in \mathcal{P}(\Theta)} J_n(\vec{v}, P) = \lim_{M \to \infty} \inf_{P \in \mathcal{P}_M(\Theta)} J_n(\vec{v}, P) = \lim_{M \to \infty} J_n(\vec{v}, \hat{P}_M).$$

As the final term in the equation above is the composition of measurable functions, it is measurable, and thus $J(\vec{v}, \hat{P}(\vec{v}))$ must be measurable, so that $\hat{P}(\vec{v})$ must be measurable as well. $\square$

4.2 Consistency of the Estimator

Theorem 4.1 shows that for any fixed n the estimator P_n and the corresponding estimate $\hat{P}_n$ exist as measurable functions mapping the data into the space of probability measures. An obvious question then, is what the resulting measures P_n or $\hat{P}_n$ represent. Since $\hat{P}_n$ is just a realization of P_n (given a specific set of data), we focus on characterization of the properties of the estimator P_n. Given the problem formulation (1.5) and the statistical model (1.3), one would certainly hope that the estimator provides some information regarding the underlying 'true' distribution P_0. In particular, we would hope that $P_n \to P_0$ in some appropriate sense. If this is the case, then the estimator is said to be *consistent*. Of course, the estimator itself is a random variable, and thus this convergence must be discussed in terms of probability. With this in mind, we consider the following set of assumptions.

(A1) For any fixed n, the error random variables $\{\mathcal{E}_j\}_{j=1}^n$ are independent and identically distributed, defined on some probability triple $(\Omega, \Sigma_\Omega, P_\Omega)$.

(A2) For $\vec{\mathcal{E}} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)$, $E[\vec{\mathcal{E}}] = 0$ and $Var[\vec{\mathcal{E}}] = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix.

(A3) (Θ, d) is a separable, compact metric space; the space $\mathcal{P}(\Theta)$ is taken with the Prohorov metric ρ.

(A4) For all j, $1 \leq j \leq n$, $t_j \in T$ for some compact space T.

(A5) The model function $v \in C(\mathcal{P}(\Theta), C(T))$.

(A6) There exists a measure μ on T such that

$$\frac{1}{n}\sum_{j=1}^{n} g(t_j) = \int_T g(t) d\mu_n(t) \to \int_T g(t) d\mu(t)$$

for all $g \in C(T)$.

(A7) The functional

$$J_0(P) = \sigma^2 + \int_T \left(v(t; P_0) - v(t; P)\right)^2 d\mu(t)$$

is uniquely minimized at $P_0 \in \mathcal{P}(\Theta)$.

Assumption (A1) establishes the probability triple on which the error random variables $\mathcal{E}_j$ are assumed to be defined. As we will see, this probability triple will permit us to make probabilistic statements regarding the consistency of the estimator P_n. These assumptions as well as the two theorems below follow closely the theoretical results of [15] which establish the consistency of the ordinary least squares estimator for a traditional nonlinear least squares problem. The key idea is to first argue that the functions $J_n(\vec{V};P)$ converge to J_0 as n increases; then the minimizer P_n of J_n should converge to the unique minimizer P_0 of J_0 [1].

Because the functions J_n are functions of the vector $\vec{V}$, which itself depends on the random variables $\mathcal{E}_j$, these functions are themselves random variables, as are the estimators P_n. Though we have generally refrained from doing so up to this point, it will occasionally be convenient to evaluate these functions at points in the underlying probability triple. Thus we may write $J_n(\vec{V};P)(\omega)$, $\mathcal{E}_j(\omega)$, etc., whenever the particular value of ω is of interest.

Theorem 4.2. *Under assumptions (A1)-(A7), there exists a set $A \in \Sigma_\Omega$ with $P_\Omega(A) = 1$ such that for all $\omega \in A$,*

$$\frac{1}{n} J_n(\vec{V};P)(\omega) \to J_0(P)(\omega)$$

as $n \to \infty$ and for each $P \in \mathcal{P}(\Theta)$. Moreover, the convergence is uniform on $\mathcal{P}(\Theta)$.

Proof. As in [15], the proof will proceed in three parts. First, for any fixed element $P \in \mathcal{P}(\Theta)$, a set A_P is constructed with $P_\Omega(A_P) = 1$ such that the convergence statement holds. The sets A_P are then used to construct a set A as described. Finally, the uniform convergence is shown.

Let $P \in \mathcal{P}(\Theta)$ be fixed. We may rewrite

$$\begin{aligned}
\frac{1}{n} J_n(\vec{V};P) &= \frac{1}{n}\sum_{j=1}^{n} \left(V_j - v(t_j;P)\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n} \left(\mathcal{E}_j + v(t_j;P_0) - v(t_j;P)\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathcal{E}_j^2 + \frac{2}{n}\sum_{j=1}^{n} \left(v(t_j;P_0) - v(t_j;P)\right)\mathcal{E}_j + \frac{1}{n}\sum_{j=1}^{n} \left(v(t_j;P_0) - v(t_j;P)\right)^2 .
\end{aligned}$$

We consider the three terms on the right. For the first term, define

$$B_1 = \left\{ \omega \in \Omega \middle| \frac{1}{n}\sum_{j=1}^{n} \mathcal{E}_j(\omega)^2 \to \sigma^2 \right\}.$$

By the Strong Law of Large Numbers, $P_\Omega(B_1) = 1$. For the third term, observe that

$$\frac{1}{n}\sum_{j=1}^{n} \left(v(t_j;P_0) - v(t_j;P)\right)^2 \to \int_T \left(v(t;P_0) - v(t;P)\right)^2 d\mu(t) = J_0(P) - \sigma^2$$

by assumption (A6) and the continuity of $v(t;\cdot)$. (Note also that this convergence is independent of $\omega \in \Omega$.) For the second term, define

$$\tilde{\mathcal{E}}_j = \left(v(t_j;P_0) - v(t_j;P)\right)\mathcal{E}_j.$$

Then

$$\begin{aligned} E[\tilde{\mathcal{E}}_j] &= 0 \\ Var[\tilde{\mathcal{E}}_j] &= \sigma^2 \left(v(t_j;P_0) - v(t_j;P)\right)^2 \\ &\leq \sigma^2 \sup_{t\in T} \left(v(t;P_0) - v(t;P)\right)^2 \\ &\leq M_P \end{aligned}$$

where the final inequality follows from the continuity of v and the compactness of T. Hence we have

$$\sum_{j=1}^{\infty} \frac{Var[\tilde{\mathcal{E}}_j]}{j^2} \leq M_P \sum_{j=1}^{\infty} \frac{1}{j^2} < \infty$$

and therefore the set B_P defined by

$$B_P = \left\{ \omega \in \Omega \Big| \frac{2}{n} \sum_{j=1}^{n} \left(v(t_j;P_0) - v(t_j;P)\right) \mathcal{E}_j \to 0 \right\}$$

satisfies $P_\Omega(B_P) = 1$ by Kolmogorov's Law of Large Numbers. Finally, we may define $A_P = B_1 \cap B_P$. Then $P_\Omega(A_P) = 1$ and $\frac{1}{n} J_n(\vec{V};P)(\omega) \to J_0(P)$ for each $\omega \in A_P$, which completes the first part of the proof.

For the second part of the proof, we must find a set A with $P_\Omega(A) = 1$ such that $\frac{1}{n} J_n(\vec{V};P)(\omega) \to J_0(P)$ for each $\omega \in A$ and for all $P \in \mathcal{P}(\Theta)$. Naively, we desire $A = \cap A_P$, but this intersection is (in general) uncountable. Rather, we construct the set A using the dense countable subset of $\mathcal{P}(\Theta)$ (Theorem 3.9). Define

$$A_1 = \left\{ \omega \Big| \frac{1}{n} \sum_{j=1}^{n} |\mathcal{E}_j(\omega)| \to E[|\mathcal{E}_1(\omega)|] \right\}.$$

Again by the Strong Law of Large Numbers, $P_\Omega(A_1) = 1$. Now define the set $\tilde{\mathcal{P}}_d(\Theta)$ as before and set

$$A = A_1 \bigcap \left[\bigcap_{P\in\tilde{\mathcal{P}}_d} A_P \right].$$

Since the intersection is taken over a countable number of sets, each having probability one (with respect to P_Ω), $P_\Omega(A) = 1$. To complete the second part of the proof, we must show that $A \subset A_P$ for all $P \in \mathcal{P}(\Theta)$ (and not merely for all $P \in \tilde{\mathcal{P}}_d(\Theta)$, which holds by the definition of A). If this is the case, then $\frac{1}{n} J(\vec{V};P)(\omega) \to J_0(P)$ for all ω in A and for all $P \in \mathcal{P}(\Theta)$.

Consider any $P \in \mathcal{P}(\Theta)$ and take $\omega \in A$, $\epsilon > 0$. Since $\omega \in A$, $\omega \in A_1$ and we may choose n_1 such that for all $n \geq n_1$,

$$\frac{1}{n} \sum_{j=1}^{n} |\mathcal{E}_j| < 1 + E[|\mathcal{E}_1|].$$

By the continuity of v and the density of $\tilde{\mathcal{P}}_d(\Theta)$ in $\mathcal{P}(\Theta)$, we may choose $P_M \in \tilde{\mathcal{P}}_d(\Theta)$ such that

$$\sup_{t\in T} |v(t;P) - v(t;P_M)| < \frac{\epsilon}{4\,(E[|\mathcal{E}_1|]+1)}.$$

Finally, $\omega \in A$ implies $\omega \in A_{P_M}$ which in turn implies $\omega \in B_{P_M}$. Thus we may choose n_2 such that for all $n \geq n_2$,

$$\left| \frac{2}{n} \sum_{j=1}^{n} \left(v(t_j;P_0) - v(t_j;P_M)\right) \mathcal{E}_j \right| < \frac{\epsilon}{2}.$$

Then for $n \geq \max\{n_1, n_2\}$,

$$\begin{aligned} \left| \frac{1}{n} J_n(\vec{V};P) - J_0(P) \right| &\leq \left| \sigma^2 - \frac{1}{n} \sum_{j=1}^{n} \mathcal{E}_j^2 \right| + \left| \frac{2}{n} \sum_{j=1}^{n} \left(v(t_j;P_0) - v(t_j;P)\right) \mathcal{E}_j \right| \\ &\quad + \left| \frac{1}{n} \sum_{j=1}^{n} \left(v(t_j;P_0) - v(t_j;P)\right)^2 - \int_T \left(v(t;P_0) - v(t;P)\right)^2 d\mu(t) \right| \end{aligned}$$

The first term goes to zero since $\omega \in A$ implies $\omega \in B_1$. The final term goes to zero by assumptions (A5) and (A6). For the second term,

$$\begin{aligned} \left| \frac{2}{n} \sum_{j=1}^{n} \left(v(t_j;P_0) - v(t_j;P)\right) \mathcal{E}_j \right| &\leq \left| \frac{2}{n} \sum_{j=1}^{n} \left(v(t_j;P_0) - v(t_j,P_M)\right) \mathcal{E}_j \right| + \frac{2}{n} \sum_{j=1}^{n} |v(t_j;P_M) - v(t_j;P)| \cdot |\mathcal{E}_j| \\ &< \frac{\epsilon}{2} + \left(2 \sup_{t\in T} |v(t;P_M) - v(t;P)| \right) \left(\frac{1}{n} \sum_{j=1}^{n} |\mathcal{E}_j| \right) \\ &< \frac{\epsilon}{2} + 2 \left(\frac{2}{4\,(E[|\mathcal{E}_1|]+1)} \right) (E[|\mathcal{E}_1|]+1) \\ &< \epsilon. \end{aligned}$$

Thus $\frac{1}{n} J_n(\vec{V};P)(\omega) \to J_0(P)$ and thus $\omega \in A_P$. Thus $A \subset A_P$ for all $P \in \mathcal{P}(\Theta)$ and the second part of the proof is complete.

Finally, we must show the convergence is uniform on $\mathcal{P}(\Theta)$ for $\omega \in A$. To do so we will show that the sequence of functions $\frac{1}{n} J_n(\vec{V};P)(\omega)$ is equicontinuous (viewed as functions of P) and then use the Arzela-Ascoli Theorem. For fixed $\omega \in A$, let $\epsilon > 0$. Take $P \in \mathcal{P}(\Theta)$. By the continuity of v (A5) and compactness of T (A4), there exists a $\delta > 0$ such that

$$\begin{aligned} \sup_{t\in T} \left| v(t;P) - v(t;\tilde{P}) \right| &< \frac{1}{6} \left\{ \frac{\epsilon}{E[|\mathcal{E}_1|]+1}, \frac{\epsilon}{\sup_{t\in T} |v(t;P_0)|} \right\} \\ \sup_{t\in T} \left| v(t;P)^2 - v(t;\tilde{P})^2 \right| &< \frac{\epsilon}{3}, \end{aligned}$$

for all $\tilde{P} \in B_\delta(P)$. Since $\omega \in A$, $\omega \in A_1$ and we can choose N such that $n \geq N$ implies

$$\frac{1}{n} \sum_{j=1}^{n} |\mathcal{E}_j| < E[|\mathcal{E}_1|] + 1.$$

Then for $n \geq N$ and for all $\tilde{P} \in B_\delta(P)$,

$$\begin{aligned}
\left|\frac{1}{n}J_n(\vec{V};P) - \frac{1}{n}J_n(\tilde{P})\right| &\leq \left|\frac{1}{n}\sum_{j=1}^{n}\left(\mathcal{E}_j + v(t_j;P_0) - v(t_j;P)\right)^2 - \frac{1}{n}\sum_{j=1}^{n}\left(\mathcal{E}_j + v(t_j;P_0) - v(t_j;\tilde{P})\right)^2\right| \\
&= \left|\frac{1}{n}\sum_{j=1}^{n}\left(2\mathcal{E}_j + v(t_j;P_0) - v(t_j;P) - v(t_j;\tilde{P})\right)\left(v(t_j;\tilde{P}) - v(t_j;P)\right)\right| \\
&\leq \left|\frac{2}{n}\sum_{j=1}^{n}\left(\mathcal{E}_j + v(t_j;P_0)\right)\left(v(t_j;\tilde{P}) - v(t_j;P)\right)\right| \\
&\quad + \frac{1}{n}\sum_{j=1}^{n}\left|v(t_j;P)^2 - v(t_j;\tilde{P})^2\right| \\
&\leq \frac{2}{n}\sum_{j=1}^{n}|\mathcal{E}_j|\left(\sup_{t\in T}\left|v(t;P) - v(t;\tilde{P})\right|\right) \\
&\quad + \sum_{j=1}^{n}\frac{2}{n}\left(\sup_{t\in T}|v(t;P_0)|\right)\left(\sup_{t\in T}|v(t;P) - v(t;\tilde{P})|\right) \\
&\quad + \sup_{t\in T}\left|v(t;P)^2 - v(t;\tilde{P})^2\right| \\
&\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.
\end{aligned}$$

Thus the sequence of functions $\frac{1}{n}J_n(\vec{V};P)(\omega)$ is equicontinuous for each $\omega \in A$ and by the Arzela-Ascoli Theorem, $\frac{1}{n}J_n(\vec{V};P)(\omega) \to J_0(P)$ uniformly on compact subsets of $\mathcal{P}(\Theta)$, and hence on $\mathcal{P}(\Theta)$ itself. □

Theorem 4.3. *Under assumptions (A1)-(A7), the estimators* $P_n \xrightarrow{w^*} P_0$ *as* $n \to \infty$ *with probability 1. That is,*

$$P_\Omega\left(\left\{\omega \middle| P_n(\vec{V})(\omega) \to P_0\right\}\right) = 1.$$

Proof. Take the set A as in the previous theorem and fix $\omega \in A$. Then by the previous theorem, $\frac{1}{n}J_n(\vec{V};P)(\omega) \to J_0(P)$ for all $P \in \mathcal{P}(\Theta)$. Let $\delta > 0$ be arbitrary and define $O = B_\delta(P_0)$. Then O is open in $\mathcal{P}(\Theta)$ (in the subspace topology) and O^C is compact (again, in the subspace topology). Since P_0 is the unique minimizer of $J_0(P)$ by assumption (A7), there exists $\epsilon > 0$ such that

$$J_0(P) - J_0(P_0) > \epsilon$$

for all $P \in O^C$. By the previous theorem, there exists n_0 such that for $n \geq n_0$,

$$\left|\frac{1}{n}J_n(\vec{V};P)(\omega) - J_0(P)\right| < \frac{\epsilon}{4}$$

for all $P \in \mathcal{P}(\Theta)$. Then for $n \geq n_0$ and $P \in O^C$,

$$\begin{aligned}\frac{1}{n}\left(J_n(\vec{V};P)(\omega) - J_n(\vec{V};P_0)(\omega)\right) &= \frac{1}{n}J_n(\vec{V};P)(\omega) - J_0(P) + J_0(P) - J_0(P_0) + J_0(P_0) \\ &\quad - \frac{1}{n}J_n(\vec{V};P_0)(\omega) \\ &\geq -\frac{\epsilon}{4} + \epsilon - \frac{\epsilon}{4} > 0.\end{aligned}$$

But $J_n(\vec{V};P_n)(\omega) \leq J_n(\vec{V};P_0)(\omega)$ by definition of P_n. Hence we must have $P_n \in O = B_\delta(P_0)$ for all $n \geq n_0$, which implies $P_n(\omega) \xrightarrow{w^*} P_0$ since $\delta > 0$ was arbitrary. □

Theorem 4.3 establishes the consistency of the estimator (1.5). Given a set of data $\vec{v}$, it follows that the estimate $\hat{P}_n$ corresponding to the estimator P_n will converge to the true distribution P_0 under the stated assumptions. We remark that these assumptions are not overly restrictive (compare [15, 20, 28]) though some of the assumptions may be difficult to verify in practice. Assumptions (A3)–(A5) are mathematical in nature and may be verified directly for each specific problem. Assumptions (A1) and (A2) describe the error process which is assumed to generate the collected data. While it is unlikely that one will be able to prove a priori that the error process satisfies these assumptions, posterior analysis such as residual plots [24, Ch. 3] can be used to investigate the appropriateness of the assumptions of the statistical model. Assumption (A6) reflects the manner in which data is sampled and, together with Assumption (A7), constitutes an identifiability condition for the model. The limiting sampling distribution function μ may be known if the experimenter has complete control over the values t_j of the independent variables (e.g., if the t_j are measurement times) but this is not always the case.

5 Computational Convergence

To this point, the analysis has focused on the properties of the estimators P_n and the resulting estimates $\hat{P}_n$. However, it is generally not possible to solve the optimization problems (1.5) or (1.6) for P_n or $\hat{P}_n$ as a function of $\vec{V}$ or $\vec{v}$. As a result, approximate (generally numerical) methods must be used in order to solve (1.7) and obtain an approximate estimate $\hat{P}^N_{n,M}$. We must ascertain, then, how the approximate estimate $\hat{P}^N_{n,M}$ relates to the exact estimate $\hat{P}_n$ (for any fixed value of n.) These results are outlined in [21] and are included again here with proof.

Theorem 5.1. *Let (Θ, d) be a compact, separable metric space and consider the space $(\mathcal{P}(\Theta), \rho)$ of probability measures on Θ with the Prohorov metric, as before. Let $\mathcal{P}_M(\Theta)$ be as defined in (4.1). Assume*

1. *the map $P \mapsto J^N_n(\vec{v}, P)$ is continuous for all n, N;*

2. *for any sequence of probability measures $P_k \to P$ in $\mathcal{P}(\Theta)$, $v^N(t;P_k) \to v(t;P)$ as $N, k \to \infty$;*

3. *$v(t;P)$ is uniformly bounded for all t, P.*

Then there exists minimizers $\hat{P}^N_{n,M}$ satisfying (1.7)*. Moreover, for fixed n, there exists a subsequence (as $M, N \to \infty$) of the approximate estimates $\hat{P}^N_{n,M}$ which converges to some (possibly non-unique) $\hat{P}^*_n$ which satisfies* (1.6).

Proof. For any fixed n, the existence of the minimizers $\hat{P}^N_{n,M}$ follows from the compactness of the space $(\mathcal{P}(\Theta), \rho)$ (Corollary 3.8) and the continuity of the map $P \mapsto J(\vec{v}; P)$ (Assumption 1). By definition, these minimizers satisfy

$$J^N_n(\vec{v}, \hat{P}^N_{n,M}) \le J^N_n(\vec{v}, P_M) \tag{5.1}$$

for all $P_M \in \mathcal{P}_M(\Theta)$ and for each n, N.

Next, we show an auxiliary result. Consider any sequence $P_k \to P$ in $\mathcal{P}(\Theta)$ (see Assumption 2). Then

$$\begin{aligned}\left|J^N_n(\vec{v}, P_k) - J_n(\vec{v}, P)\right| &= \left|\sum_j \left(v_j - v^N(t_j; P_k)\right)^2 - \sum_j \left(v_j - v(t_j; P)\right)^2\right| \\ &= \left|\sum_j \left(2v_j - v^N(t_j; P_k) - v(t_j; P)\right)\left(v(t_j; P) - v^N(t_j; P_k)\right)\right| \\ &< M \sum_j \left|v(t_j; P) - v^N(t_j; P_k)\right| \to 0,\end{aligned}$$

where we have used the uniform boundedness of $v(t; P)$ (Assumption 3) as well as Assumption 2.

Now, we return to (5.1). Since $\mathcal{P}(\Theta)$ is compact there must exist (possibly after reindexing) a limit $\hat{P}^*_N = \lim \hat{P}^N_{n,M}$. Next consider any $P \in \mathcal{P}(\Theta)$. By Theorem 3.9, it is possible to construct a sequence of measures $P_M \in \mathcal{P}_M(\Theta) \subset \mathcal{P}(\Theta)$ so that $P_M \to P$ in $\mathcal{P}(\Theta)$. Hence, taking limits in (5.1) as M and N go to infinity (where we tacitly assume the indices have been renumbered according to the convergent subsequence), we have

$$J_n(\hat{P}^*_n) \le J_n(P) \text{ for all } P \in \mathcal{P}(\Theta),$$

and we see that $\hat{P}^*_n$ satisfies (1.6). □

This theorem provides a set of conditions under which a subsequence of approximate estimates $\hat{P}^N_{n,M}$ converges to the estimate $\hat{P}^*_n$ of interest. This estimate is itself a realization (for a particular data set) of the estimator P_n which has been shown to exist and to be consistent, so that $P_n \to P_0$ with probability one. Thus we have some reasonable assurance that a computed approximate estimate $\hat{P}^N_{n,M}$ reflects the true distribution P_0. The assumptions of Theorem 5.1 are not restrictive. In typical problems (and, indeed, in the assumptions of other theorems appearing in this document) it is assumed that the parameter space Θ as well as the independent variable space T are compact (see, e.g., Section 4). In such a case, Assumptions 1 and 3 above are satisfied if the individual model solutions $y(t; \theta)$ are continuous on $T \times \Theta$. Assumption 2 is then simply a condition on the convergence of the numerical procedure used in obtaining model solutions.

Significantly, the Prohorov Metric Framework is computationally constructive. In practice, one does not construct a sequence of estimates for increasing values of M and N; rather, one fixes the values of M and N to be sufficiently large to attain a desired level of accuracy. By Theorem 3.9, we need only to have some enumeration of the elements of $\mathcal{P}_M(\Theta)$ in order to compute an approximate estimate $\hat{P}^N_{n,M}$. (We will not consider the choice of N, as this will depend upon the numerical framework by which approximate model solutions $v^N(t;P)$ are obtained.) Practically, this is accomplished by selecting M nodes in Θ, $\{\theta^M_k\}^M_{k=1}$. The optimization problem (1.7) is then reduced to a standard constrained estimation problem over Euclidean M-space in which one determines the values of the weights p^M_k corresponding to each node. Thus,

$$\begin{aligned}\hat{P}^N_{n,M} &= \arg\min_{\mathcal{P}_M(\Theta)} \sum_{j=1}^{n} \left(v_j - v(t_j;P)\right)^2 \\ &= \arg\min_{\mathcal{P}_M(\Theta)} \sum_{j=1}^{n} \left(v_j - \int_\Theta Cy(t_j;\theta)dP(\theta)\right)^2 \\ &= \arg\min_{\widetilde{\mathbb{R}^M}} \sum_{j=1}^{n} \left(v_j - \left(\sum_{k=1}^{M} Cy(t_j;\theta^M_k)p^M_k\right)\right)^2,\end{aligned}$$

where in the final line we seek the weights $\bar{p}^M = (p^M_1, \ldots, p^M_M)^T \in \widetilde{\mathbb{R}^M} = \{\bar{p}^M | p^M_k \in \mathbb{R}^+, \sum_{k=1}^M p^M_k = 1\}$. These are sufficient to characterize the approximating discrete estimate $\hat{P}^N_{n,M}$ since the nodes are assumed to be fixed in advance. Moreover, define

$$\begin{aligned}H_{kl} &= 2\sum_j \left(Cy(t_j;\theta_k)\right)\left(Cy(t_j;\theta_l)\right) \\ f_k &= -2\sum_j v_j \left(Cy(t_j;\theta_k)\right) \\ c &= \sum_j \left(v_j\right)^2.\end{aligned}$$

Then one can equivalently compute [11]

$$\hat{P}^N_{n,M} = \arg\min_{\widetilde{\mathbb{R}^M}} \left(\frac{1}{2}\left(\bar{p}^M\right)^T H\bar{p}^M + f^T\bar{p}^M + c\right). \tag{5.2}$$

From this reformulation, it is clear that the approximate problem (1.7) has a unique solution if H is positive definite. If the individual mathematical model (1.2) is independent of P,[2] then the

[2] This independence of the individual model on the population distribution is strongly suggested by our choice of notation for the individual solutions, $y(t;\theta,\psi)$. In many problems of interest, this is a perfectly reasonable assumption. For instance, in a size-structured biological model [9, 11, 13, 16, 17], the individual rate of growth may vary across the population, but the rate of growth of an individual is unaffected by the rates of growth of his neighbors. It is possible however, that the individual mathematical model may depend upon the population distribution, $y(t;\theta,\psi,P)$. For instance, in a size-structured population model, fast-growing individuals may out-compete their slower growing neighbors for limited resources. Such examples also arise in models of electromagnetic polarization and deformations of viscoelastic materials. See [2, Sec. 14.1.2] for a more complete discussion.

matrices H and f can be precomputed in advance. Then one can rapidly (and exactly) compute the gradient and Hessian of the objective function in a numerical optimization routine. As M grows large, the quadratic optimization problem (5.2) becomes poorly conditioned [11]. Thus there is a trade-off: M must be chosen sufficiently large so that the computational approximation is accurate, but not so large that ill-conditioning leads to large numerical errors. The efficient choice of M as well as the choice of the nodes $\{\theta_k\}_{k=1}^{M}$ is an open research problem.

It should be acknowledged that the uniqueness of the computational problem (i.e., when H is positive definite) is not sufficient to ensure the uniqueness of the limiting estimate $\hat{P}_n^*$ in Theorem (5.1) (as there could be multiple convergent subsequences). However, if $J_n(\vec{v}; P)$ is uniquely minimized, then every subsequence of $\hat{P}_{n,M}^N$ which converges as N, M grow large must converge to that unique minimizer. Moreover, under assumptions (A1)–(A7) in Section 4, it has been shown that $\frac{1}{n}J_n(\vec{v}, P) \to J_0(P)$ (as n grows large) with probability one, and the function $J_0(P)$ is assumed to be uniquely minimized by P_0.

6 Extensions to Other Error Models

In this final section, we comment on generalizations of the statistical and error models (1.3) and (1.4). As noted in Section 1, the estimator (1.5) is premised upon an assumption of independent, identically distributed, constant variance additive error,

$$V_j = v(t; P_0) + \mathcal{E}_j,$$

which may be rewritten

$$\vec{V} = v(\vec{t}; P_0) + \vec{\mathcal{E}} \tag{6.1}$$

where by assumption

$$\begin{aligned} E[\vec{\mathcal{E}}] &= \vec{0} \\ Var[\vec{\mathcal{E}}] &= \sigma^2 I_n. \end{aligned}$$

While such an assumption is common, many physical and biological problems are not accurately described by such a simple statistical model. Thankfully, the results presented above can be easily extended to cover a larger class of error models. Consider the more general error model

$$\begin{aligned} E[\vec{\mathcal{E}}] &= \vec{0} \\ Var[\vec{\mathcal{E}}] &= \sigma^2 W = \sigma^2 diag(w(t_1)^2, \ldots, w(t_n)^2), \end{aligned} \tag{6.2}$$

where the function $w(t) > 0$ is a continuous weighting function. Such a statistical model arises from an observation process in which measurement errors are independent but are not identically distributed. This formulation includes the special case that $w(t) = v(t; P_0)$, which is commonly called a *relative error* [24] or *constant coefficient of variance* (CCV) error model [20, 26]. (Of course, in such a case, one does not actually know P_0 and an iterative estimation procedure must be used [24, 26].) Now define $L = diag(w(t_1), \ldots, w(t_n))$. It follows that $LL^T = L^2 = W$ and L^{-1} exists (since it is assumed $w(t) > 0$ for all t). Applying L^{-1} to both sides of (6.1),

$$L^{-1}\vec{V} = L^{-1}\vec{v}(\vec{t}, P_0) + L^{-1}\vec{\mathcal{E}}$$

or

$$\vec{Z} = \vec{\nu}(\vec{t}, P_0) + \vec{\eta}, \tag{6.3}$$

where $\vec{Z}$, $\vec{\nu}$, and $\vec{\eta}$ have the obvious definitions. Moreover, assuming the distributions from which the random errors are drawn are uniquely determined by their first two statistical moments, the random variables η_j are independent and identically distributed with constant variance. Thus the theory presented in this document can be applied to the transformed model (6.3).

Additional generalizations are also possible. For instance, the matrix W may depend upon additional nuisance parameters γ. In particular it has been shown that histogram data from a flow cytometer is well-described by an error model of the form $W = diag(w(t_1)^\gamma, \ldots, w(t_n)^\gamma)$ for some scalar γ [20, 23, 32]. Such nuisance parameters can be estimated in an iterative procedure [28] and the theory presented in this document is essentially unchanged. If the observations are not independent, then the matrix R in (1.4) will not be diagonal. In such a situation, the theory presented in this report can still be applied provided R is diagonalizable and this diagonalization is sufficient so that the resulting transformed errors are independent and identically distributed (such is the case, for instance, for autoregressive errors of order $r < n$).

7 Concluding Remarks

In this document we have defined a parameter estimation problem in which one has a mathematical model describing the dynamics of an individual biological or physical process but data which is sampled from a population of individuals. Because each individual is assumed to be described by a unique set of parameters, the data is described not by a single parameter but by the probability distribution (over all individuals) from which these individual parameters are sampled. Theoretic results for the nonparametric measure estimation problem are presented which establish the existence and consistency of the estimator. A previously proposed and numerically tested computational scheme is also discussed and its convergence is proven.

Several open problems remain. First, while the computational scheme is simple, it is not always clear how one should go about choosing the M nodes θ_k from the dense subset of Θ which are then used to estimate weights p_k. From a theoretical perspective, the nodes need only to be added so that they ‘fill up’ the parameter space in an appropriate way. In practice, however, rounding error and ill-conditioning can be quite problematic, particularly for a poor choice of nodes. A more complete computational algorithm would include information on how to optimally choose the M nodes θ_k (as well as the appropriate values of M).

Additionally, given the consistency of the estimator P_n, it would be desirable to place some measure of confidence on the estimated probability distribution. The traditional frequentist approach relies on either asymptotic theory or bootstrapping to construct such measures of confidence. In the former case, it is not clear how one might extend notions of sensitivity to the space of probability measures, which would require a notion of differentiability on the space of probability measures. In the latter case, the results provide some computational estimates but a rigorous theory is not yet available. Some preliminary work on these topics has been initiated [12, 14] and is still ongoing.

8 Acknowledgements

This research was supported in part by grant number NIAID R01AI071915-09 from the National Institute of Allergy and Infectious Diseases and in part by the U.S. Air Force Office of Scientific Research under grant number AFOSR FA9550-12-1-0188.

References

[1] Takeshi Amemiya, Nonlinear regression models, Ch. 6 in *Handbook of Econometrics, Volume I*, Z. Griliches and M. D. Intriligator, Eds. North Holland, Amsterdam, 333–389.

[2] H.T. Banks, *A Functional Analysis Framework for Modeling, Estimation, and Control in Science and Engineering*, CRSC Press, Boca Raton, 2012.

[3] H.T. Banks, J.E. Banks and S.L. Joyner, Estimation in time-delay modeling of insecticide-induced mortality, CRSC-TR08-15, North Carolina State University, October 2008; *J. Inverse and Ill-posed Problems*, **17** (2009), 101–125.

[4] H.T. Banks and Kathleen Bihari, Modelling and estimating uncertainty in parameter estimation, *Inverse Problems*, **17** (2001), 95–111.

[5] H. T.Banks, V. A. Bokil, S. Hu, F. C. T. Allnutt, R. Bullis, A. K. Dhar and C. L. Browdy, Shrimp biomass and viral infection for production of biological countermeasures, CRSC-TR05-45, December, 2005; *Mathematical Biosciences and Engineering*, **3** (2006), 635–660.

[6] H.T. Banks and D.M. Bortz, Inverse problems for a class of measure dependent dynamical systems, *J. Inverse and Ill-posed Problems*, **13** (2005), 103–121.

[7] H.T. Banks, D.M. Bortz and S.E. Holte, Incorporation of variability into the mathematical modeling of viral delays in HIV infection dynamics, *Mathematical Biosciences*, **183** (2003), 63–91.

[8] H.T. Banks, D.M. Bortz, G.A. Pinter and L.K. Potter, Modeling and imaging techniques with potential for application in bioterrorism, CRSC-TR03-02, North Carolina State University, January 2003; Chapter 6 in *Bioterrorism: Mathematical Modeling Applications in Homeland Security* (H.T. Banks and C. Castillo-Chavez, eds.), Frontiers in Applied Math, **FR28**, SIAM, Philadelphia, 2003, 129–154.

[9] H.T. Banks, L.W. Botsford, F. Kappel and C. Wang, Modeling and estimation in size structured population models, LCDS-CCS Report 87-13, Brown University; *Proc. 2nd Course on Mathematical Ecology*, (Trieste, December 8–12, 1986) World Press (1988), Singapore, 521–541.

[10] H. T. Banks, M. W. Buksas, T. Lin. *Electromagnetic Material Interrogation Using Conductive Interfaces and Acoustic Wavefronts.* SIAM FR **21**, Philadelphia, 2002.

[11] H.T. Banks and J.L. Davis, A comparison of approximation methods for the estimation of probability distributions on parameters, *Appl. Num. Math.*, **57** (2007), 753–777.

[12] H.T. Banks and J.L. Davis, Quantifying uncertainty in the estimation of probability distributions, *Math. Biosci. Eng.,* **5** (2008), 647–667.

[13] H.T. Banks, J.L. Davis, S.L. Ernsterberger, S. Hu, E. Artimovich and A.K. Dhar, Experimental design and estimation of growth rate distributions in size-structured shrimp populations, CRSC-TR08-20, North Carolina State University, November 2008; *Inverse Problems* **25** (2009), 095003 (28 pages).

[14] H.T. Banks, S. Dediu and H.K. Nguyen, Sensitivity of dynamical systems to parameters in a convex subset of a topological vector space, *Math. Biosci. Eng.,* **4** (2007), 403–430.

[15] H.T. Banks and B.G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, *J. Math. Biol.,* **28** (1990), 501–527.

[16] H.T. Banks and B.G. Fitzpatrick, Estimation of growth rate distributions in size structured population models, *Quarterly of Applied Mathematics*, **49** (1991), 215–235.

[17] H.T. Banks, B.G. Fitzpatrick, L.K. Potter and Y. Zhang, Estimation of probability distributions for individual parameters using aggregate population data, CRSC-TR98-06, North Carolina State University, January 1998; in *Stochastic Analysis, Control, Optimization, and Applications*, (W. McEneaney, G. Yin, and Q. Zhang, eds.), Birkhauser, Boston, 1989.

[18] H.T. Banks and N.L. Gibson, Well-posedness in Maxwell systems with distributions of polarization relaxation parameters, CRSC-TR04-01, North Carolina State University, January 2004; *Appl. Math. Letters,* **18** (2005), 423–430.

[19] H. T. Banks and N. L. Gibson, Electromagnetic inverse problems involving distributions of dielectric mechanisms and parameters, CRSC-TR05-29, August, 2005; *Quarterly of Applied Mathematics*, **64** (2006), 749–795.

[20] H.T. Banks, Z.R. Kenz and W.C. Thompson, An extension of RSS-based model comparison tests for weighted least squares, *Intl. J. Pure and Appl. Math,* **79** (2012), 155–183.

[21] H.T. Banks, Z.R. Kenz and W.C. Thompson, A review of selected techniques in inverse problem nonparametric probability distribution estimation, CRSC-TR12-13, North Carolina State University, May 2012; *J. Inverse and Ill-Posed Problems*, to appear.

[22] H. T. Banks and G. A. Pinter, A probabilistic multiscale approach to hysteresis in shear wave propagation in biotissue, CRSC-TR04-03, January, 2004; *SIAM J. Multiscale Modeling and Simulation*, **3** (2005), 395–412.

[23] H.T. Banks, W.C. Thompson, C. Peligero, J. Argilaguet and A. Meyerhans, Experimental and biological variability in CFSE-based flow cytometry data, *in preparation.*

[24] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton, London, New York, 2009.

[25] P. Billingsley, *Convergence of Probability Measures*, Wiley & Sons, New York, 1968.

[26] M. Davidian and D.M. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, London, 2000.

[27] O. van Gaans, Probability measures on metric spaces, *Available Online*, <http://www.math.leidenuniv.nl/∼vangaans/jancol1.pdf>

[28] A. Ronald Gallant, *Nonlinear Statistical Models*, John Wiley and Sons, New York, 1987.

[29] Yu. V. Prohorov, Convergence of random processes and limit theorems in probability theory, *Theor. Prob. Appl.,* **1** (1956), 157–214.

[30] H.L. Royden, *Real Analysis*, 3rd Ed., Prentice Hall, Upper Saddle River, New Jersey, 1988.

[31] G.A. Seber and C.J. Wild, *Nonlinear Regression*, Wiley, Hoboken, 2003.

[32] W. Clayton Thompson, *Partial Differential Equation Modeling of Flow Cytometry Data from CFSE-based Proliferation Assays,* Ph.D. Dissertation, Dept. of Mathematics, North Carolina State University, Raleigh, December, 2011.

[33] W. Whitt, *Stochastic-Process Limits*, Springer-Verlag, New York, 2002.